# **Project Report for**

# **Machine Learning Group Project**

# 

## **Master’s in Data Science and Advanced Analytics at NOVA IMS, Lisbon**

## **Group Information**

Group: MAA\_202021\_50

Group members:

Md. Shawkatul Islam Aziz

Md. Tahir Hossain

Philipp Metzger

Ali Sabbir

Henrique Vaz

**Abstract**

((Summarise the introduction))

This project’s goal is to implement Machine Learning methods and algorithms for classifying individuals as having an income below or above average. For this, with a training dataset that was provided, multiple techniques are implemented and tested.

((Summarise the methodology))

In order to complete the task an exploratory analysis of the training dataset is conducted, several transformations are applied to certain variables, the existence of missing values is assessed, the discriminatory power of categorical features is analysed, the categorical variables are encoded, an outlier detection is performed, the features are scaled, and an appropriate subset of the original features is selected for further processing. Then, ((…))

((Summarise the Results))

((Summarise the Conclusions))

**I. Introduction**

The goal of this project is to implement Machine Learning algorithms that given the suitable input data are capable of predicting if an individual has an income lower or higher than the average income in a group of citizens.

For this, a dataset of 22400 observations serves as training data. This dataset includes amongst others the variables ‘Birthday’, ‘Marital Status’ and ‘Education Level’ and the binary target variable ‘Income’. A full list of the variables contained in the dataset will be presented in Chapter III. The input data and target vector are used to train several predictive models of which one is chosen in the end to be the best suited model for the task.

The project is related to the Kaggle competition ‘Newland’ in which several groups of students of the course Data Science and Advanced Analytics at the Lisbon based university NOVA IMS compete in a competition that is embedded in the fictitious scenario of the colonisation of a newly discovered habitable planet. In this Kaggle competition, each group uploads a vector of predictions computed with their best model, based on the input data of a test dataset provided in the project materials. The vector of predictions serves as the quality measure to assess which group designed and implemented the best predictive model.

**II. Background**

In this chapter, the theoretical background of the techniques or algorithms who have not been explored during the practical classes but are applied in this project are explained.

Support Vector Classification:

“Support vector machine” (SVM) belongs to the supervised machine learning algorithms. It can be used for regression and classification. The latter is the case in this project. Thus, the algorithm used here is called support vector classifier. The goal of support vector machines is to define a hyperplane in the n-dimensional space containing the training data points[[1]](#footnote-1), that separates the different classes in the target variable[[2]](#footnote-2). In case of a linearly separable dataset, meaning that a hyperplane can be defined such that all observations from one class are on one side and all observations from the other class are on the other side of the hyperplane, the hyperplane simply represents the border that separates the two classes and that has the maximum distance to each and every point in the dataset (measured individually), which means that only the distances to the points on each side that are closest to the hyperplane are relevant. These closest points are referred to as the “support vectors” since they support the position of the hyperplane. Because the area around the hyperplane could be imagined as a channel or street separating the dataset, this is also called the “widest street approach”. In case of a non-linearly separable dataset, a penalty can be added during the calculation of the optimal solution, such that points that are “on the wrong side” of the hyperplane are penalized in a way that is adequate for the application. This is called the “Soft Margin” approach (see chapter 7.1.1 in [3]). In applications, where the dataset contains patterns where a hyperplane is not able to create a satisfactory division, such as when a ring of points of one class surrounds a cluster of points of the other class, the n-dimensional space of the datapoints can be mapped to a higher-dimensional space in order to achieve a better separability. This is called the “Kernel Trick” (see chapter 6 in [3]).

Random Oversampling:

When one group of a categorical or binary target variable has more observations than the others (in the categorical case) or the other (in the binary case), it can be sensible to use random oversampling to harmonise the number of observations associated to each category of the variable. In random oversampling this is achieved by duplicating observations that are associated to the category or the categories that originally have less observations associated to them. The observations to duplicate are selected randomly and with repetition. The goal of random oversampling can be that after the process, all categories have the same number of observations associated with them. Or in the binary case, it is also be an option to define a factor, by which the numbers of observations of the two classes differ from each other such that n\_small = a \* n\_large, ∈ IR, a ∈ (0,1], where n\_small is the number of observations associated with the class that originally had less observations and n\_large is the number of observations associated with the other class. Random oversampling is described in Chapter 5 of [1].

**III. Methodology** ((check the order of everything in the notebook))

**III.1 Materials and Software**

The materials used in this project are the training dataset and the test dataset provided. The original training dataset consists of 22400 observations and 15 variables (including the target variable). These variables are shown in table 1 which is from the file ‘Project Presentation.pdf’ which serves for defining the fictitious background and the goal of the project.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Citizen\_ID | Unique identifier of the citizen |
| Name | Name of the citizen (First name and surname) |
| Birthday | The date of Birth |
| Native Continent | Home continent of the citizen on planet Earth |
| Marital Status | The marital status of the citizen |
| Lives with | The household environment of the citizen |
| Base Area | The neighborhood of the citizen in Newland |
| Education Level | The education level of the citizen |
| Years of Education | The number of years of education of the citizen |
| Employment Sector | The employment sector of the citizen |
| Role | The job role of the citizen |
| Working Hours per week | The number of working hours per week of the citizen |
| Money Received | The money payed to the elements of Group B |
| Ticket Price | The money received by the elements of Group C |
| Income | The dependent variable (Where 1 is Income higher than the average and 0 Income Lower or equal to the average) |

Table 1: The original dataset’s variables adopted from “Project Presentation.pdf”

The software that is used in order to complete the project is Python and more precisely Anaconda and Jupyter Notebook. In the latter, the code for this project is created.

In the following part, the steps conducted in our Jupyter notebook are described.

**III.2 Loading of Data, Data Exploration and Data Pre-processing**

First, all packages and libraries that are used are imported. These include, amongst others, libraries from the standard Python library such as ‘os’ and packages from ‘sklearn’ which are used for the predictive models but also the library ‘xgboost’ which we used for feature selection, but not for classification.

In the second step the data described above is loaded into main memory and a first data exploration is conducted through which it becomes clear, that the training dataset contains more observations with the target variable ‘Income’ equal to 0 than observations with the target variable ‘Income’ equal to 1, meaning that in the training dataset, there are more observations that are classified as having an income lower than the average.[[3]](#footnote-3)

In the following step the variable ‘Birthday’ is transformed to obtain the age of each individual. As the unit for this newly created variable days is chosen, and the old variable is replaced by the new one.

In the next step, the existence of cells containing empty strings, spaces or NaN values[[4]](#footnote-4) is assessed.

In a further exploratory analysis, the discriminatory power of the categorical features in the training dataset is assessed by plotting bar charts for each category in which the affiliation the either target value is represented in either grey or green.

The following step’s purpose is to one-hot encode most of these categorical features in order to make it possible to process them in the algorithms used later on. Only the feature ‘Education Level’ is not one hot encoded. For this one, a different encoding method is designed and applied, which assigns a numerical rank to each original value of this variable.

Next, outlier detection is performed. For this, the minima and maxima of features that could potentially have outliers are computed and assessment is made whether or not these values are realistic. The result of this assessment is described in chapter IV.

The following step is the procedure of feature scaling. For this, a standard scaler is used, which standardises every feature by removing its mean and scaling it to unit variance [2].

**III.3 Feature Selection**

Next, as a first feature selection step, correlations between the metric variables as well as between all variables are assessed to decide, which features are contributing only redundant information to the dataset. The outcome of this analysis is described in more detail in chapter IV; for now it is relevant to know that some features are discarded and a list of features ‘features\_to\_keep\_1’ is defined, which contains the features that are kept after the correlation analysis. The variables of the resulting training dataset are presented in table 2.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Age\_days\_rel\_to\_2020 | Age of the citizen relative to the most recent day the code has been run |
| Marital\_Status\_Married | 1: Citizen is married, 0 otherwise |
| Marital\_Status\_Single | 1: Citizen is single, 0 otherwise |
| Marital\_Status\_Divorced | 1: Citizen is divorced, 0 otherwise |
| Lives\_with\_Children | 1: Citizen lives with children, 0 otherwise |
| Lives\_with\_Husband | 1: Citizen lives with husband, 0 otherwise |
| Lives\_with\_Alone | 1: Citizen lives alone, 0 otherwise |
| Lives\_with\_Other Family | 1: Citizen lives with other family, 0 otherwise |
| Role\_Management | 1: Citizen’s role is in management, 0 otherwise |
| Role\_Other services | 1: Citizen’s role is in “other services”, 0 otherwise |
| Role\_? | 1: Citizen’s role is unknown, 0 otherwise |
| Role\_Administratives | 1: Citizen’s role is of administrative nature, 0 otherwise |
| Role\_Cleaners & Handlers | 1: Citizen’s role is described as “Cleaners & Handlers”, 0 otherwise |
| Role\_Professor | 1: Citizen is a professor, 0 otherwise |
| Working Hours per week | The number of working hours per week of the citizen |
| Money Received | The money payed to the elements of Group B |
| Ticket Price | The money received by the elements of Group C |
| Education\_Level\_Classified | The education level of the citizen in respect to the classification implemented in this project |
| Employment\_Sector\_Self-Employed (Company) | 1: Self-employed (Company), 0 otherwise |
| Employment\_Sector\_Private Sector - Services | 1: Employed in private sector, 0 otherwise |
| Income | The dependent variable (Where 1 is Income higher than the average and 0 Income Lower or equal to the average) |

Table 2: The training dataset’s variables after the pre-processing steps

In a further feature selection step, RFE, a Ridge classifier and XGBoost are applied in order to identify the most important features. The result of this further limitation of the feature space is saved in the list ‘features\_to\_keep\_2’. As it is explained in Chapter IV, in the course of the project, the decision was made to use the features in ‘features\_to\_keep\_1’ for all models and to disregard the limitations of the feature space suggested by this step.

The goal of the next steps is to assess the best feature combination for an MLP[[5]](#footnote-5) classifier that is used for this project. This assessment was conducted using the classifier with its default parameters on varying subsets of the training dataset presented in table 2. For evaluating the performance of the MLP classifier on the varying datasets, k-fold cross validation with k = 10 was used.

Next, the same type of analysis was conducted for another classification algorithm: An AdaBoost classifier[[6]](#footnote-6).

((still to write: Everything including and after Model selection))

**IV. Results**

**V. Discussion**

**VI. Conclusion**

**VII. REFERENCES**

[1] Fernández, Alberto & García, Salvador & Galar, Mikel & Prati, Ronaldo & Krawczyk, Bartosz & Herrera, Francisco. (2018). Learning from Imbalanced Data Sets. 10.1007/978-3-319-98074-4.

[2] <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>, viewed on 25. Dec. 2020 at 13:07.

[3] Bishop, Christopher. (2006). Pattern Recognition and Machine Learning. 10.1117/1.2819119.

[4] <https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html#sklearn.neural_network.MLPClassifier.score>, viewed on 25. Dec. 2020 at 16:14.

[5] <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html>, viewed on 25. Dec. 2020 at 17:39.

1. Or, if the “Kernel Trick” is used, a higher dimensional space of which said n-dimensional space is a subspace. [↑](#footnote-ref-1)
2. Typically, two classes as in this application, but adaptations for multiclass problems also exist (See chapter 7.1.3 in [3]). [↑](#footnote-ref-2)
3. Income = 1: 76.29 %, Income = 1: 23.71 % [↑](#footnote-ref-3)
4. NaN stands for Not a Number, meaning that the respective value is non-existent. [↑](#footnote-ref-4)
5. Multilayer Perceptron (for a general explanation see chapter 5 in [3]; for the documentation of the implementation used for this project see [4]) [↑](#footnote-ref-5)
6. AdaBoost, short for Adaptive Boosting (for a general explanation see chapter 14.3. in [3]; for the documentation of the implementation used for this project see [5]) [↑](#footnote-ref-6)